Summary Report On

Project US House Price Prediction (2004-2024)

Executive Summary

This project aims to predict the US House Price Index (HPI) using various economic indicators from the Federal Reserve Economic Data (FRED) database. The analysis spans from January 2004 to January 2024, employing Ridge and Lasso regression techniques to understand the relationship between house prices and economic variables. The models demonstrated strong predictive capabilities, with high R-squared values and low RMSE, indicating their effectiveness in forecasting house prices.

Introduction

Predicting house prices is crucial for various stakeholders, including policymakers, investors, and homeowners. Accurate predictions help in economic planning, investment decisions, and policy-making. This project utilizes economic indicators to build regression models that forecast the US House Price Index, providing insights into the factors influencing house prices.

Data Collection and Preparation

Data for each series was fetched using the Federal Reserve Economic Data (FRED) database API, covering the period from January 2004 to January 2024. The data was resampled to a monthly frequency to ensure uniformity across all series.

The following economic indicators were collected from the FRED database: -

- **S&P Core Logic Case-Shiller U.S. National Home Price Index (CSUSHPISA):** dependent variable representing the US house price index.
- **Interest Rates (FEDFUNDS):** The effective federal funds rate, which influences borrowing costs.

- **Unemployment Rate (UNRATE):** The percentage of the labor force that is unemployed, providing insights into economic health.
- **Consumer Price Index (CPIAUCSL):** Measures the average change in prices paid by urban consumers for goods and services, indicating inflation.
- **GDP Growth (A191RL1Q225SBEA):** Real GDP growth rate, reflecting overall economic growth.
- **Population Growth (POPTHM):** Population growth rate, influencing demand for housing.
- **Median Income (MEHOINUSA672N):** Median household income, indicating the purchasing power of consumers.
- **Consumer Confidence (UMCSENT):** University of Michigan Consumer Sentiment Index, reflecting consumer confidence in the economy.

S&P Core Logic Case-Shiller U.S. National Home Price Index (CSUSHPISA) was selected as the dependent variable (y), while the other economic indicators served as independent variables (X).

Data Cleaning

Interpolation

Missing values were filled using linear interpolation to maintain continuity in the data. This method estimates missing values by connecting the dots with a straight line, ensuring a smooth transition between data points. Linear interpolation is particularly useful for time-series data where maintaining the trend and seasonality is crucial.

Outlier Removal

Outliers were identified and removed using the Z-score method. Values beyond three standard deviations from the mean were considered outliers and excluded from the dataset. This approach helps to ensure the dataset's normality and reduce the impact of extreme values on subsequent analyses. Removing these outliers can significantly improve the accuracy and reliability of machine learning models by preventing skewed results.

Model Building

Dataset Splitting

The dataset was split into training and test sets with an 80-20 ratio to ensure a robust evaluation of model performance. The training set, comprising 80% of the data, was used to train the model, while the remaining 20% served as the test set to evaluate the model's performance. This split helps to ensure that the model is not overfitting and can generalize well to new, unseen data.

Feature Standardization

Standard-Scaler was used to standardize the feature variables, transforming them to have a mean of 0 and a standard deviation of 1. Standardization ensures that all features contribute equally to the model by bringing them onto the same scale. This step is particularly important for regularization techniques like Lasso (L1 regularization) and Ridge (L2 regularization) regression. These methods add a penalty to the loss function based on the magnitude of coefficients, and having standardized features ensures that the penalties are applied uniformly, thereby improving the stability and performance of the regression models.

Ridge Regression

Ridge regression was employed to predict the Home Price Index, with hyper parameter tuning performed using Gridn SearchCV to find the optimal alpha value.

- Best Alpha: 0.01
- Cross-Validation R^2 Scores: [0.97777552, 0.94604202, 0.93611486, 0.97216107, 0.96925206]
- Mean Cross-Validation R^2 Score: 0.9602691053198397
- Test RMSE: 8.689284255551142

- Test R-squared: 0.9730356156612795

Lasso Regression

Lasso regression was also used to predict the Home Price Index, with hyper parameter tuning performed using GridSearchCV to find the optimal alpha value.

- Best Alpha: 0.01
- Cross-Validation R^2 Scores: [0.97794676, 0.94632136, 0.93632557, 0.97195492, 0.96906544]
- Mean Cross-Validation R^2 Score: 0.9603228098954005
- Test RMSE: 8.706990223962862
- Test R-squared: 0.972925614204709

Coefficient Analysis

Ridge Regression (L2 Regularization)

The coefficients from the Ridge regression model indicate the impact of each feature on the Home Price Index while applying an L2 penalty to shrink the coefficients. This regularization technique helps to mitigate multicollinearity and prevent overfitting by constraining the coefficients, thus providing a more robust and stable model. By examining the magnitude and direction of these coefficients, we can gain insights into which features have a significant positive or negative influence on the Home Price Index.

Ridge Regression Coefficients:

Interest Rates:	-1.698010
Unemployment Rate:	-6.263968
CPI:	71.115446

GDP Growth: 3.145973

Population Growth: -55.893084

Median Income: 24.437607

Consumer Confidence: -3.426272

Lasso Regression (L1 Regularization)

The coefficients from the Lasso regression model also provide insights into the impact of each feature on the Home Price Index, but with an L1 penalty that encourages sparsity. This means that Lasso regression can effectively reduce some coefficients to zero, effectively performing feature selection. Features with non-zero coefficients in the Lasso model are identified as the most influential predictors. This not only helps in understanding the key drivers of the Home Price Index but also simplifies the model by highlighting the most important features.

By comparing the coefficients from both Ridge and Lasso models, we can identify the most consistent and impactful features and better understand the underlying factors that influence the Home Price Index.

Lasso Regression Coefficients:

Interest Rates: -1.580454

Unemployment Rate: -6.257278

CPI: 70.770587

GDP Growth: 3.136046

Population Growth: -55.472522

Median Income: 24.356269

Consumer Confidence: -3.471552

Interpretation of Coefficients

Ridge Regression

- **Interest Rates (-1.698010)**: Higher interest rates lead to lower house prices, as borrowing becomes more expensive, reducing demand for housing.
- **Unemployment Rate (-6.263968)**: Higher unemployment rates result in lower house prices due to reduced purchasing power and economic uncertainty.
- **CPI (71.115446)**: Higher consumer prices are associated with higher house prices, possibly due to inflationary effects that increase the cost of living and property values.
- **GDP Growth (3.145973)**: Higher GDP growth correlates with higher house prices, reflecting economic prosperity and increased investment in real estate.
- **Population Growth (-55.893084)**: This negative coefficient might be counterintuitive and could warrant further investigation to understand underlying factors, such as regional differences or migration patterns.
- **Median Income (24.437607)**: Higher median income leads to higher house prices due to increased purchasing power and demand for better housing.
- **Consumer Confidence (-3.426272)**: Higher consumer confidence might lead to lower house prices, which might require further analysis. This could reflect periods of economic stability where consumers feel less urgency to invest in real estate.

Lasso Regression:

- **Interest Rates (-1.580454)**: Higher interest rates lead to lower house prices, consistent with the Ridge model, reinforcing the relationship between borrowing costs and housing demand.
- **Unemployment Rate (-6.257278)**: Higher unemployment rates result in lower house prices, confirming the negative impact of economic downturns on the housing market.
- **CPI (70.770587)**: Higher consumer prices are associated with higher house prices, supporting the inflationary effect on real estate values.
- **GDP Growth (3.136046)**: Higher GDP growth correlates with higher house prices, indicating economic growth's positive influence on the housing market.

- **Population Growth (-55.472522)**: While counterintuitive, this negative coefficient may reflect regional variations or short-term economic conditions that warrant further investigation to uncover potential underlying causes, such as urbanization trends or housing supply constraints.
- **Median Income (24.356269)**: Higher median income leads to higher house prices due to increased purchasing power and demand for quality housing, consistent with the Ridge model.
- **Consumer Confidence (-3.471552)**: The negative coefficient for consumer confidence might indicate market sentiment or investment preferences during the study period, suggesting a complex relationship that requires deeper exploration. This could reflect scenarios where high confidence leads to diversified investments away from real estate.

Why Ridge Regression and Lasso Regression chosen ?

Ridge and Lasso regressions were chosen for the US House Price Prediction project because of their effectiveness in handling multicollinearity and enhancing model performance through regularization.

Handling Multicollinearity

- Ridge Regression: Adds an L2 penalty to shrink coefficients, reducing the impact of multicollinearity and making the model more stable.
- Lasso Regression: Adds an L1 penalty, shrinking some coefficients to zero, effectively performing feature selection and simplifying the model.

Regularization

- Ridge Regression: Prevents overfitting by shrinking all coefficients, ensuring better generalization to new data.
- Lasso Regression: Performs feature selection, reducing model complexity and enhancing interpretability.

Model Interpretability

- Ridge Regression: Retains all predictors with reduced impact, allowing for the examination of their relative importance.
- Lasso Regression: Identifies the most influential predictors by setting some coefficients to zero, simplifying the model.

Proven Performance

- Both models are known for their strong predictive capabilities and are reliable for forecasting tasks.

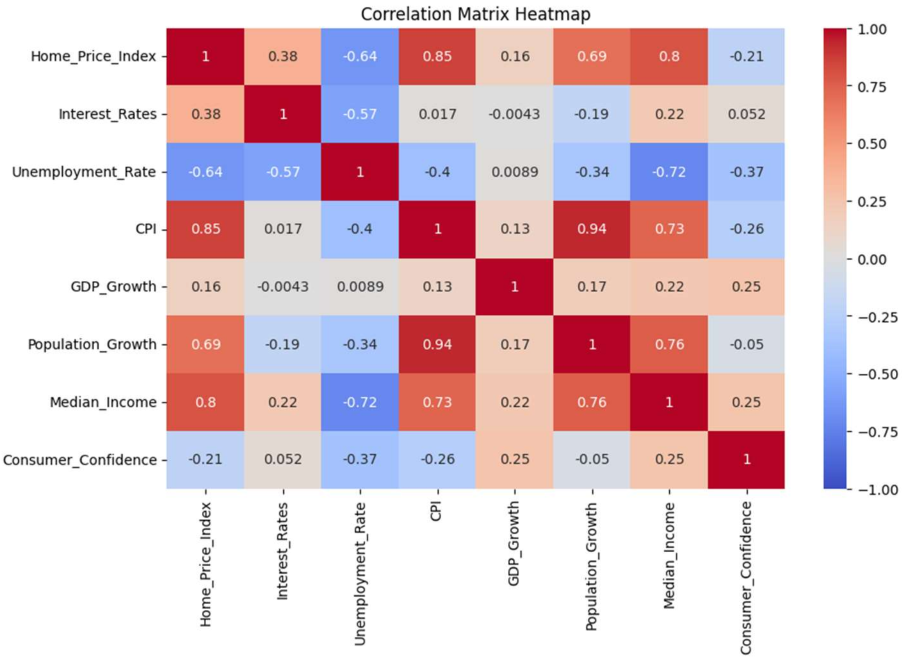
Ease of Implementation

- Both techniques can be easily implemented and tuned using common machine learning libraries and tools

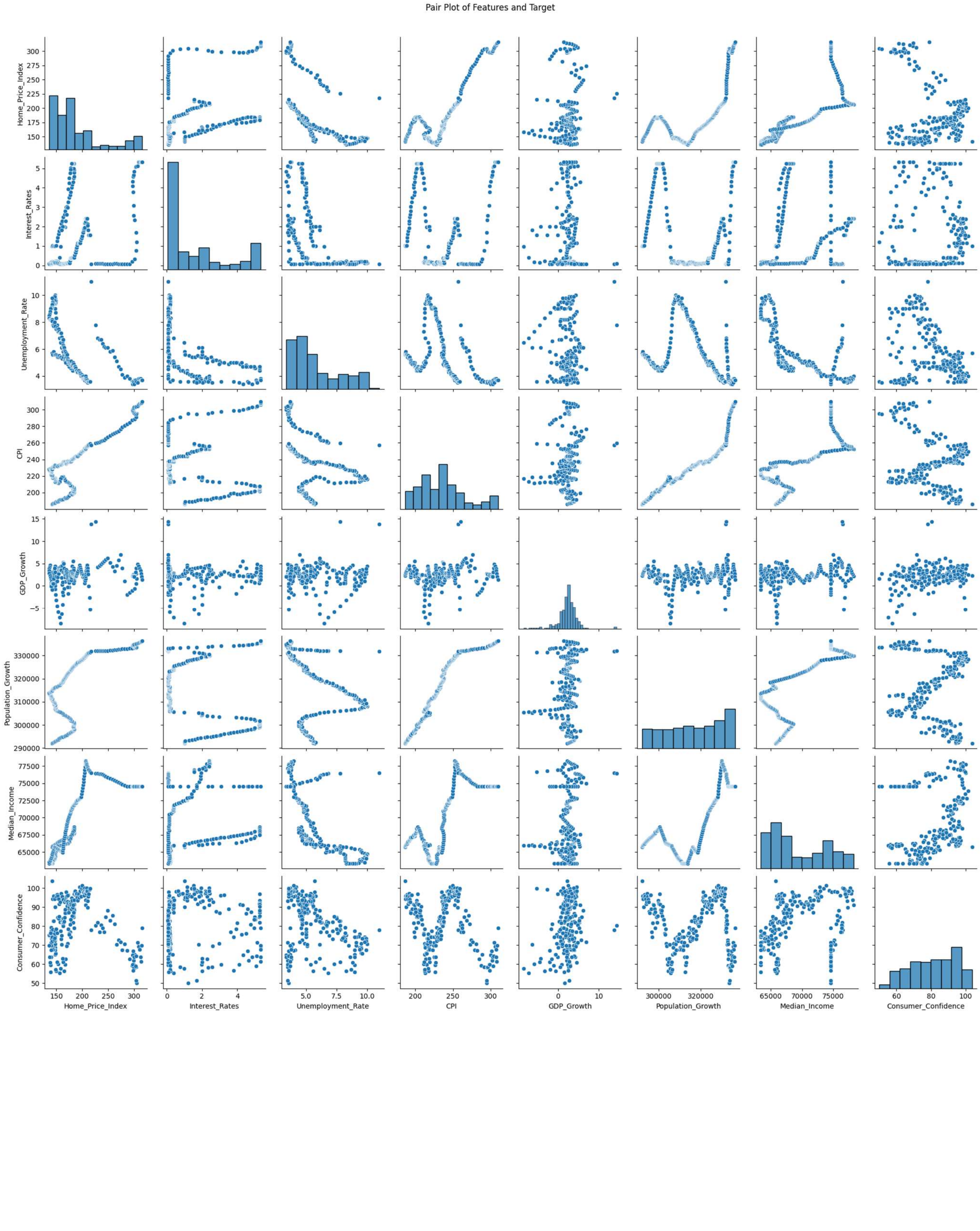
Visualizations

Various plots were generated to visualize the relationships between the independent and the dependent variable, evaluate the distribution and stability of residuals, and compare the distribution of actual vs. predicted home prices.

Correlation Heat Map

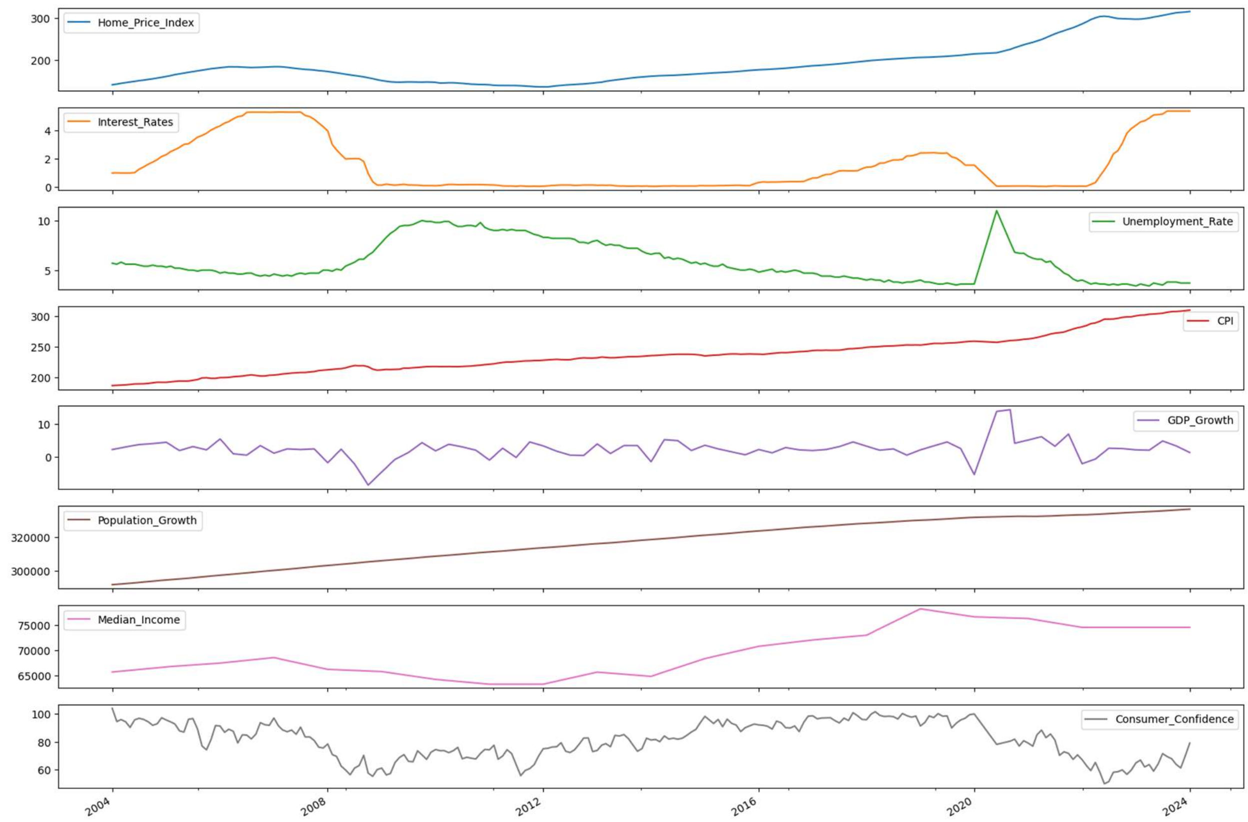


Pair Plot

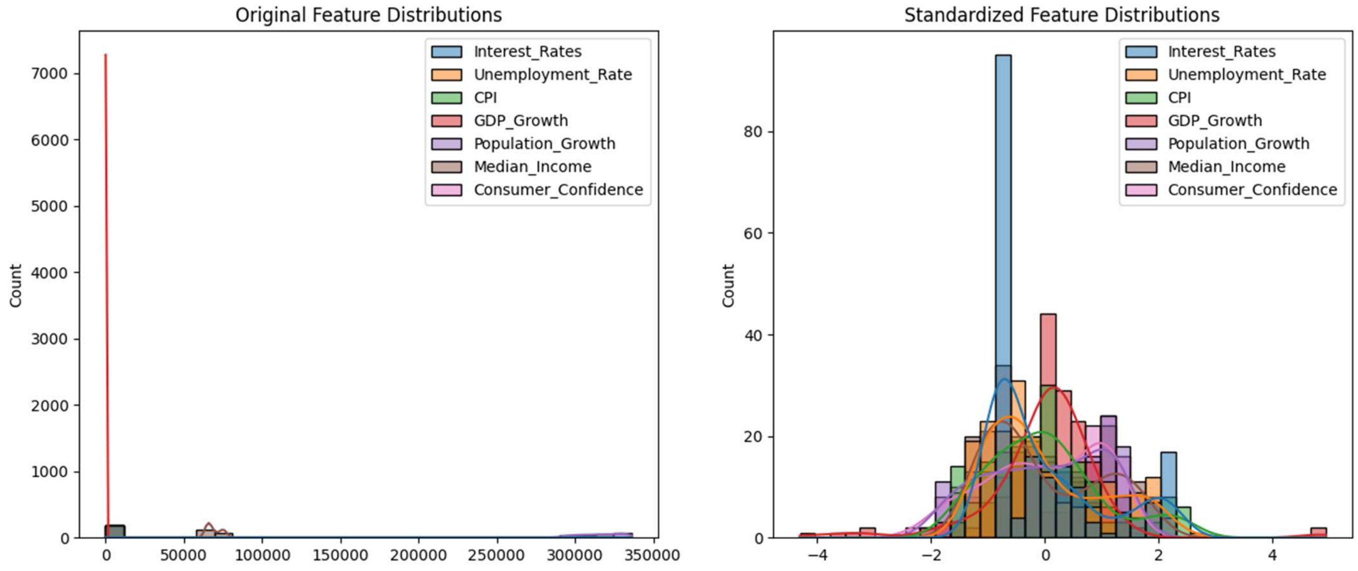


Pair Plot of Features and Target

Time Series Plot

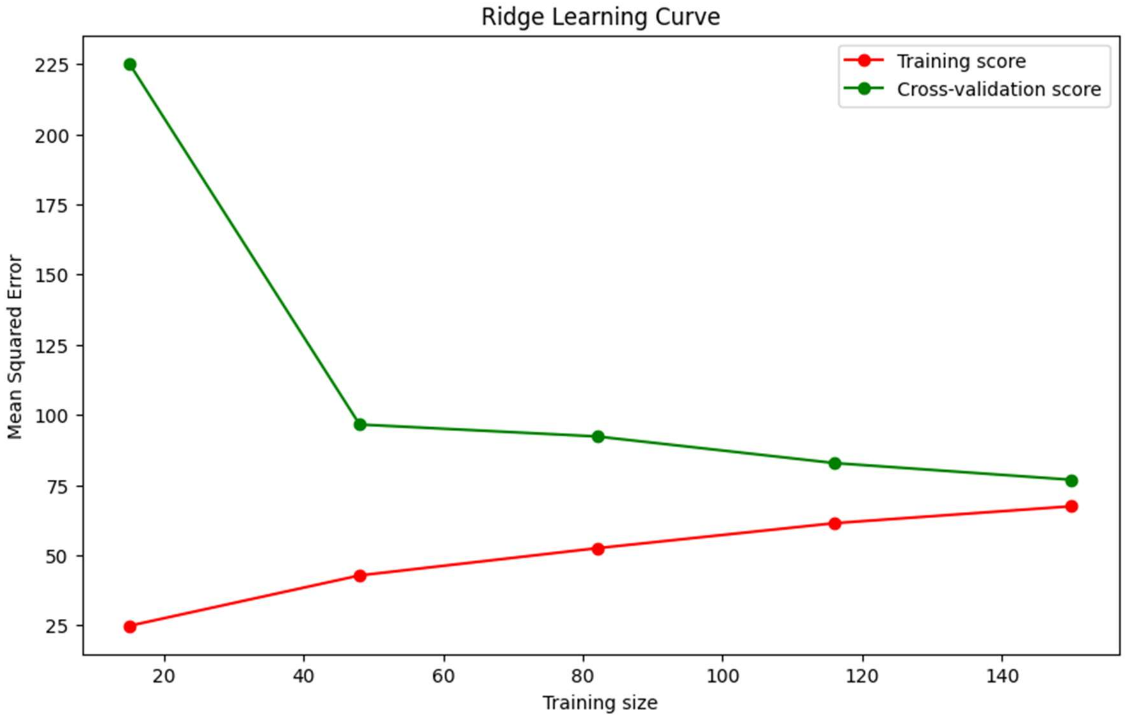


Feature Distribution Before and After Standardization

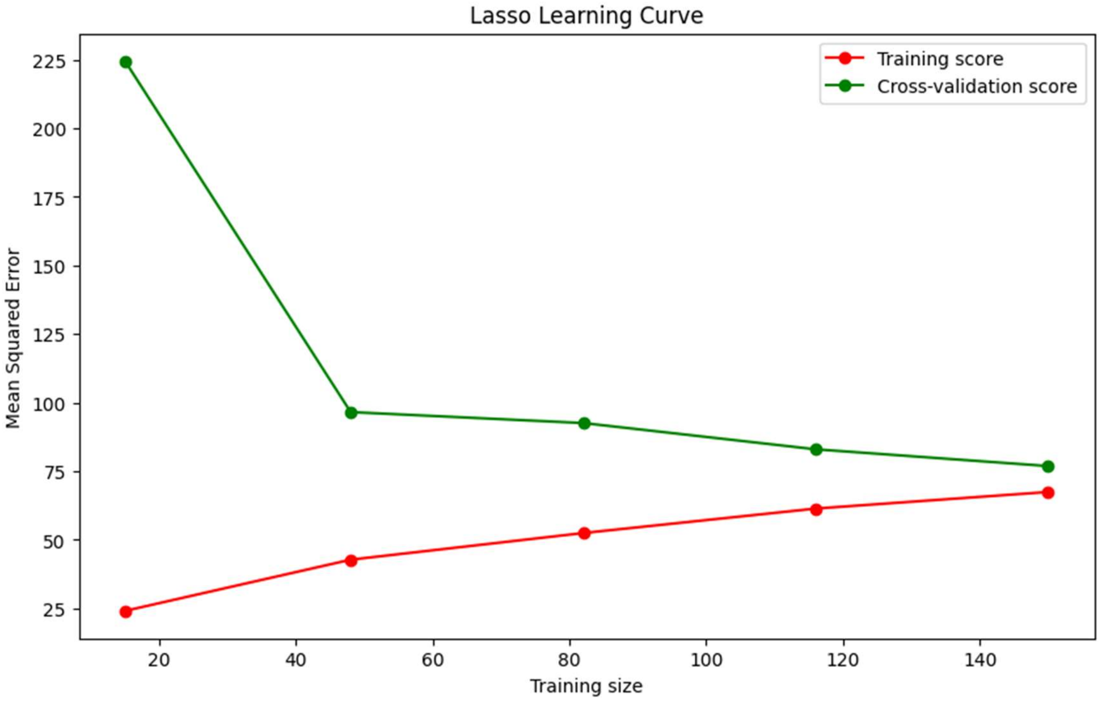


Learning Curves

Ridge Regression Learning Curve

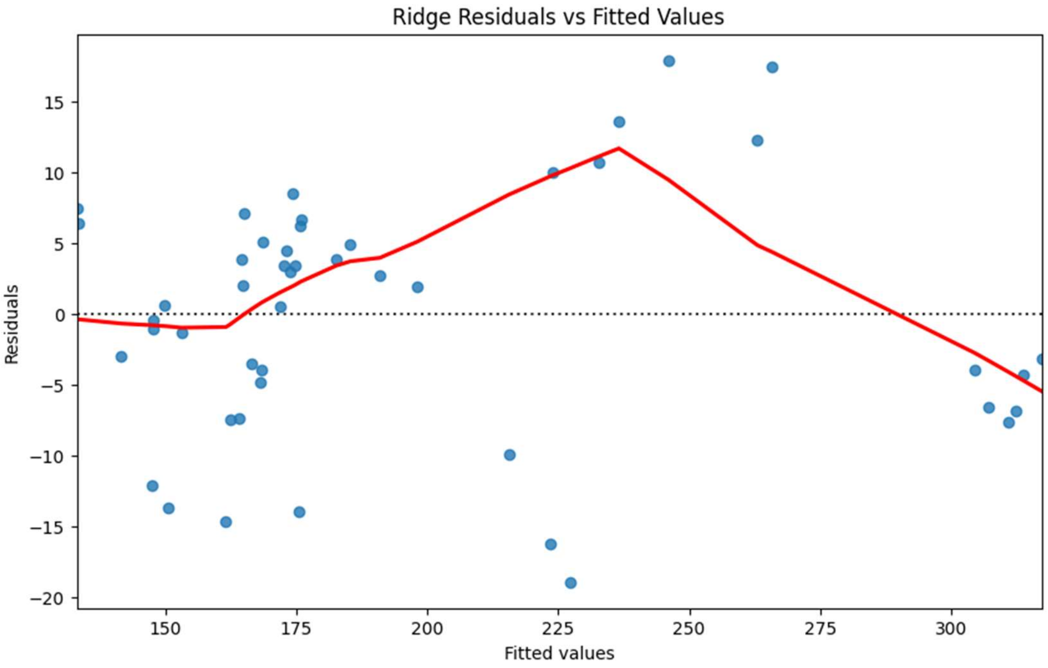


Lasso Regression Learning Curve

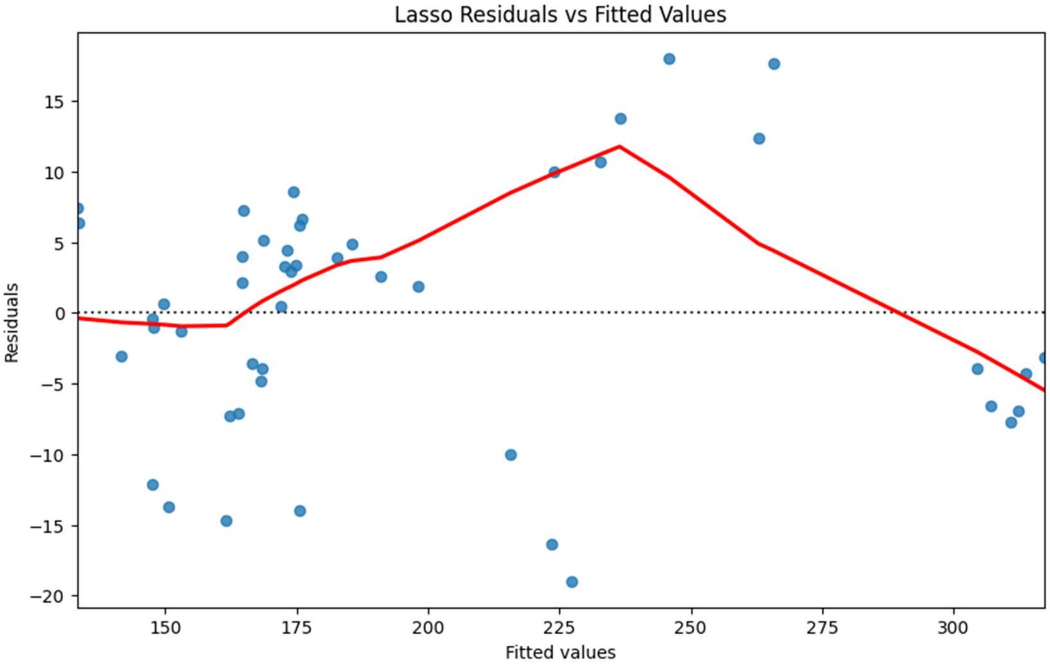


Residual Plots

Ridge Residual Vs Fitted Values

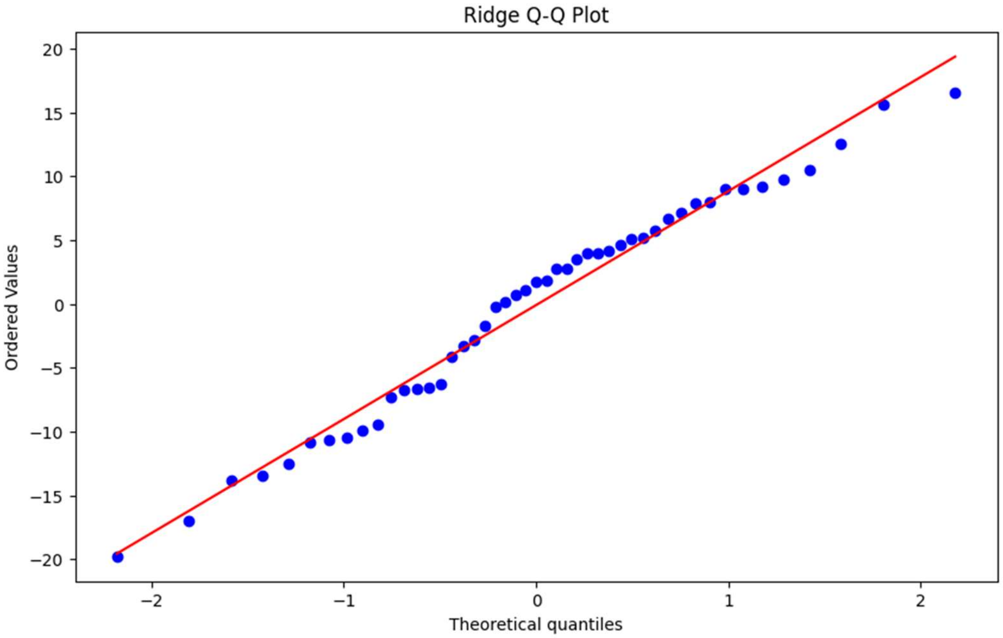


Lasso Residual Vs Fitted Values

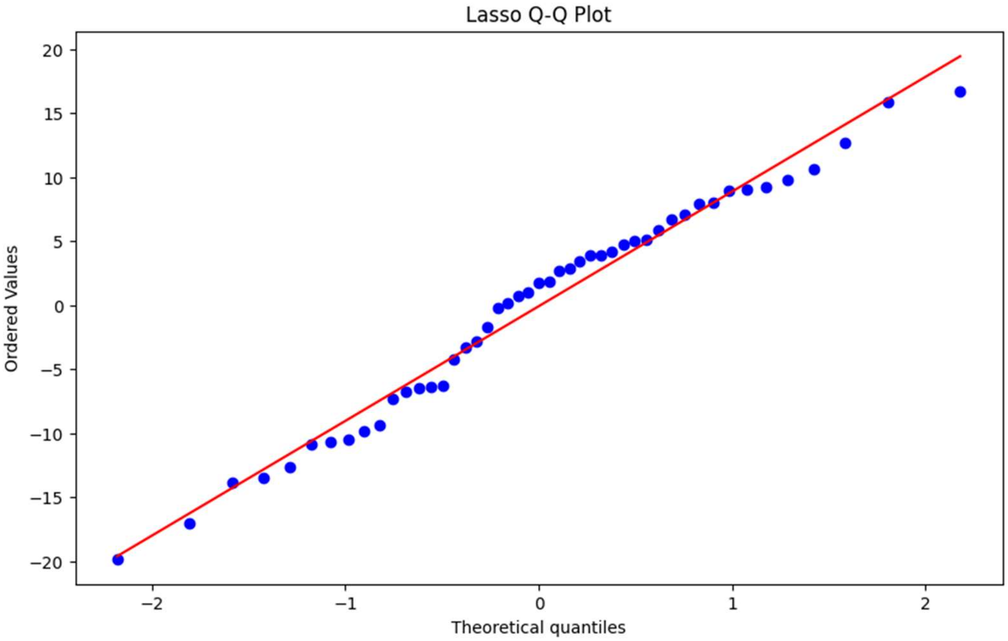


Q-Q Plots

Lasso Regression Q-Q Plot

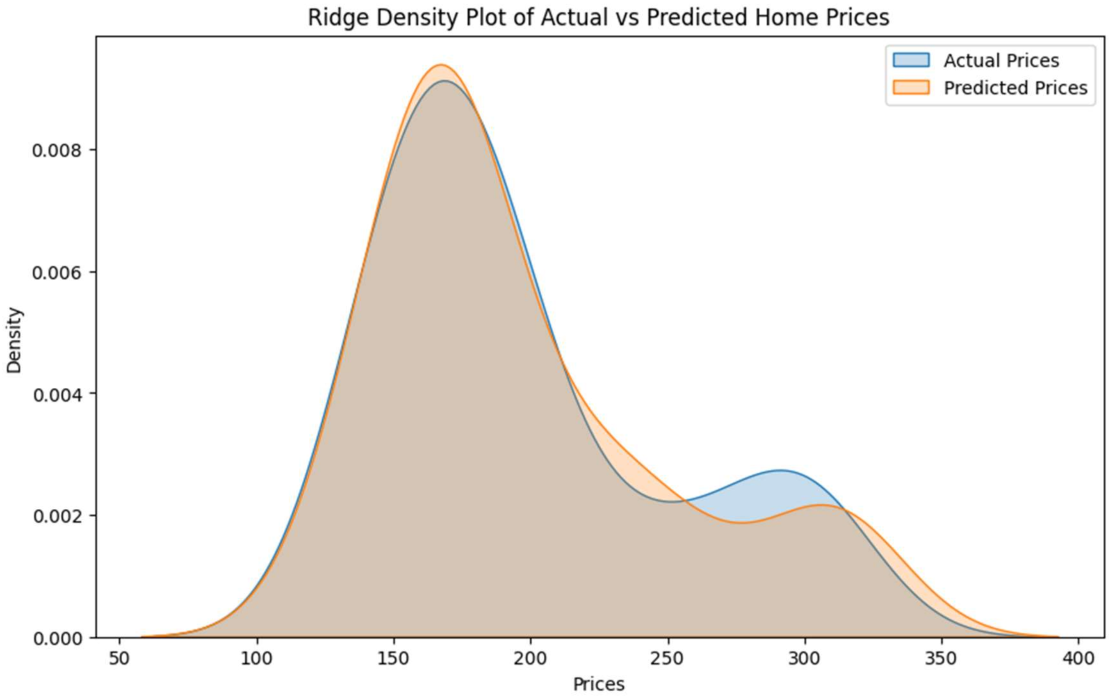


Ridge Regression Q-Q Plot

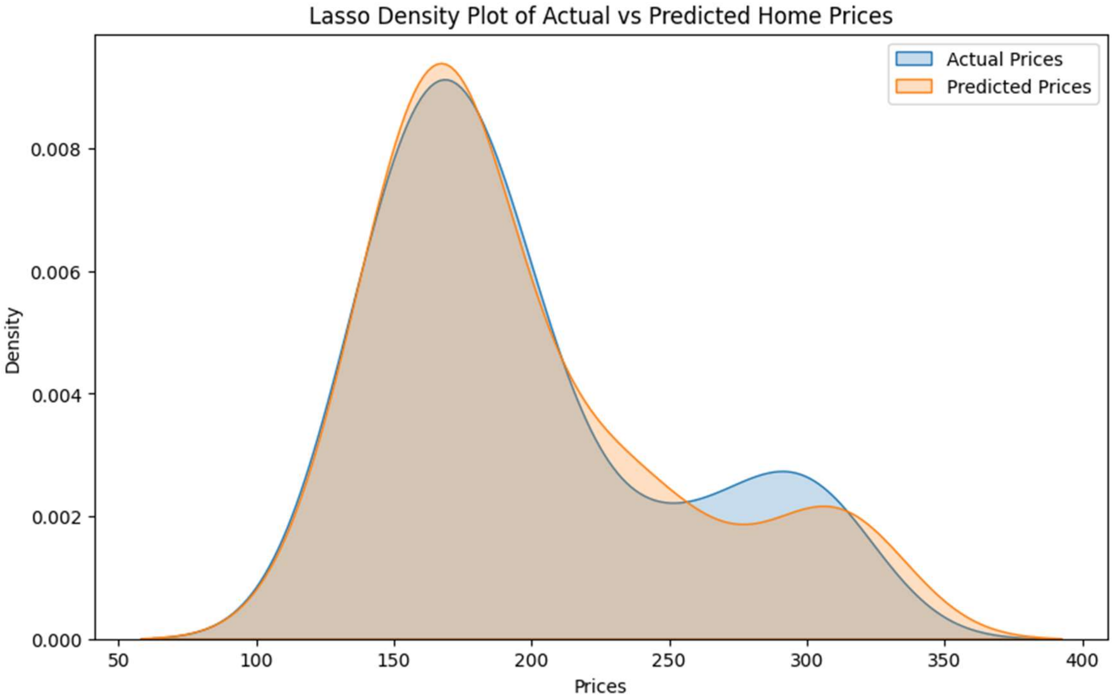


Density Plots

Actual Vs Predicted Home Prices for Ridge Regression



Actual Vs Predicted Home Prices for Lasso Regression



Use Cases

- **Policy Implications:** Policymakers can use these models to forecast housing market trends and make informed decisions. By understanding how economic indicators such as interest rates, unemployment, and GDP growth affect house prices, policymakers can design targeted interventions to stabilize the housing market and promote affordable housing initiatives.
- **Investor Insights:** Investors can leverage these predictions to make strategic investments in the housing market. By identifying trends and key drivers of house prices, investors can optimize their portfolios, manage risks, and capitalize on emerging opportunities in different regions.
- **Homeowners:** Understanding the impact of economic indicators on house prices can help homeowners make better decisions regarding buying or selling properties. With insights into market conditions, homeowners can time their transactions to maximize value.
- **Real Estate Developers:** Developers can use the model to identify lucrative locations for new projects and predict future demand. By understanding the economic factors influencing house prices, developers can make informed decisions on project.
- **Financial Institutions:** Banks and mortgage lenders can use these insights to assess the risk associated with lending. By forecasting house price trends, financial institutions can adjust their lending criteria, interest rates, and mortgage products.

Conclusion

The Ridge and Lasso regression models demonstrated strong predictive capabilities for the US House Price Index using selected economic indicators. Both models achieved high R-squared values and low RMSE values, indicating their effectiveness. The analysis of coefficients provided insights into the impact of each feature on house prices. The project highlights the importance of economic indicators in predicting housing market trends and provides a robust framework for future analyses and model improvements. Further exploration and validation using additional data and alternative modeling techniques could enhance the accuracy and robustness of the predictions.
Project Report: US House Price Prediction (2004-2024)